Advances in Cognitive Systems X (20XX) 1-6 Submitted X/20XX; published X/20XX

A Dynamic and Dual-Process Theory of Humor

Boyang Li BOYANGLI@GATECH.EDU
School of Interactive Computing, Georgia Institute of Technology, Atlanta GA 30308

Abstract

The cognitive mechanism of humor has been studied for centuries, with multiple seemingly incompatible theories proposed. Recent research in emotions suggests human emotions are tightly coupled and closely interact with other types of cognitive processes. This entangled nature contributes to the difficulty of humor research. In this paper, I attempt to provide a single, unified framework of humor, grounded in recent developments on emotion and dual-process cognition. I propose that humor comprehension consists of a four-step dynamic process: surprise, reflection, dismissal and compensation. The proposed theory provides a modern update on existing theories of humor, and is capable of explaining several phenomena that cannot be easily explained by existing theories. I also discuss the implication of the theory on creating computational systems that can create or perceive humor.

1. Introduction

Humor probably represents one of the most fluid and creative aspects of human intelligence. Popular fictions often portray Artificial Intelligence as capable but humorless. As such, the study of humor may open a door to understanding the mechanism and organization of human cognition, as well as facilitate its replication. Theories of humor date back at least to Plato's *Philebus* and Aristotle's *Poetics*, both promoting the superiority theory (e.g. Bain, 1875), which posits that we laugh at the misfortune of other people. Since then, an abundance of theories have been proposed, ranging from the release of psychic or nervous energy (Spencer, 1860; Freud, 1928; Berlyne, 1972) to the formation of an incongruity that is later resolved (e.g. Koestler, 1964; Suls, 1972; Minsky, 1980; Veatch, 1998), and so on. Each theory seems to possess some explanatory power, yet none can satisfactorily encompass all empirical evidence and provide a unified account. Reviewing these theories, one can easily be reminded of the ancient fable of blind men and the elephant.[1]

Several new theories have been proposed in the past few years (McGraw & Warren, 2010; Hurley, Dennett, & Adams, 2013; Topolinski, 2014), attempting to provide a more unified explanation for humor. A common issue with these theories is that humor is still treated as an independent affect, created by an independent cognitive sub-system and serves an independent function. In contrast, recent research suggests that emotional appraisals closely work with other cognitive sub-systems to create a rich emotional experience (Barrett, 2011; Scherer, 2001; Marsella & Gratch, 2009; Cunningham, Dunfield, & Stillman, 2013). In this paper, I provide the first attempt at explaining humor

1. Four blind men tried to figure out the shape of an elephant. They respectively touched its ear, nose, body, and leg, and all claim an elephant is just like the body part they felt.

as the interactions between cognitive sub-systems that are widely recognized to exist. As a result, I hope to develop a more parsimonious theory of humor, which hopefully also sheds light on the development of AI systems that can create and understand humor.

The theory in this paper is built on top of two main theoretical foundations: the dual-process theory (Stanovich & West, 2000; Evans, 2003; Kahneman, 2011; Evans & Stanovich, 2013), which states human cognition contains a set of automatic, effortless, fast, and intuitive processes, and a set of deliberate, effort-hungry, slow and rational processes, and theories on the dynamics of emotions and how emotions are constructed from interactions of primitive processes (Barrett, 2011; Scherer, 2001; Marsella & Gratch, 2009; Cunningham, Dunfield, & Stillman, 2013).

My main argument is that comprehension of humor is a dynamic process. It starts with a surprise that is sufficient to confuse the automatic processes and engage the deliberate processes, followed by a quick realization that the surprise is not worthy of further mental effort, which disengages the deliberate processes and produces an amplified positive emotion. This is the basic form of humor. Variations of humor are produced by different realizations. The surprising stimulus may be trivial because we discover a logic flaw, or because we see malicious intent, stupidity, or social inappropriateness. These realizations further compound the effects of humor. In this theory, the comprehension of humor is constructed from a sequence of cognitive appraisals and a quick succession of associated emotions. We no longer need to theorize a special standalone cognitive process for humor. To my best knowledge, this is the first attempt at explaining humor in terms of recent cognitive science theories.

This theory has good explanatory power as it can subsume many existing theories, including the superiority theory, the release theory, and the congruity-resolution theory. Further support comes from its ability to explain phenomena that are difficult to fit into existing theories, such as the recently reported frustration smiles and humor's persistent appeal after repetition.

In the remainder of this paper, I will first review relevant research in human cognition before introducing my theory of humor. I will compare this theory to existing theories of humor and discuss various evidences that support this theory. Finally, I will discuss the implications on the study of Artificial Intelligence aiming to reproduce human abilities with humor.

2. Theoretical Backgrounds

In this section, I introduce the two main theoretical foundations of my theory on humor: the dual-process theory, and the theories on emotion dynamics and emotion construction.

2.1 Two Types of Cognitive Processes

The dual-process theory (Stanovich & West, 2000; Evans, 2003; Evans & Stanovich, 2013; Kahneman, 2011) can be summarized as the co-existence of two types of processes in human cognition. The two types are often referred to as implicit vs. explicit, automatic vs. deliberate, etc. In this paper, I use the terminology Sytem 1 and System 2. System 1 requires little mental effort and attention, and works automatically to provide quick responses to external stimuli. System 1 is relatively inflexible and error-prone when dealing with unfamiliar problems and environments. System 2 requires substantial mental effort. It works slowly and sequentially, but is flexible enough to handle

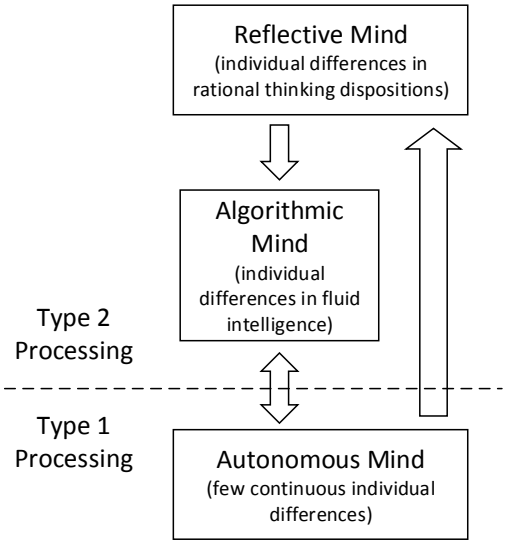

Figure 1. Stanovich's tripartite model of the two-process theory, adapted from (Evans & Stanovich, 2013).

novel and complex problems. The defining characteristic that differentiates the two systems is that System 1 does not utilize working memory, whereas System 2 requires working memory (Evans & Stanovich, 2013). As System 2 requires mental effort, we are inclined to reduce its use and delegate tasks to System 1.

The dual-process theory was first proposed to explain the individual differences in dealing with many cognitive puzzles designed to induce erroneous judgments. For example, in the famous Linda test (Tversky & Kahneman, 1983; De Neys, 2006), Linda is described with a number of features stereotypically associated with a feminist, and the test participants are asked to select the most likely from two possibilities: One, Linda is a feminist and a bank teller. Two, Linda is a bank teller. Mathematically, the second option cannot be less probable than the first option, yet most participants choose the first option. However, some participants are capable of finding the correct answer. The dual-process theory posits that the main source of individual differences exist in System 2. A tripartite model proposed by Stanovich (2011) is shown in Figure 1. The algorithmic mind represents cognitive abilities and individual differences in these abilities, such as different working memory capacity. The reflective mind reflects people's tendency to use their cognition, such as whether they think extensively or gather all evidences before making a decision.

In addition to explaining individual differences in the tastes for humor (see Section 5.4), the dual-process theory also provides insight on how cognition interacts with affective states. For example, positive emotions induce more heuristic processing (i.e. System 1), and negative emotions induce systematic processing (i.e. System 2), possibly because a negative emotion signals insufficient understanding and the need for more elaborate processing (Bagozzi, Gopinath, & Nyer, 1999; Hullett, 2005; Schwarz, 2011). Surprise is believed to provide an important signal that triggers Sys-

tem 2 (Lieberman, 2003). When our expectation is violated, the resulted indicates a novel situation that the associative System 1 is not capable of handling. For example, when we try to open a door by turning a door knob, the action is automatic as long as the door knob works as we expect. However, if turning the door knob does not open the door, deliberate processing of System 2 is needed. The engagement of System 2 by surprise is a key proposition in the present theory on humor.

2.2 Theories on the Dynamics of Emotions

The recently proposed constructionist theory of emotion (Barrett, 2006; Barrett, 2011; Cunningham, Dunfield, & Stillman, 2013) claims that emotions are not atomic, indivisible, and categorical entities. Rather, emotions are constructed by the interplay of primitive cognitive processes, many of which are not dedicated emotional processes. Emotions described by the same categorical words such as "joy" or "anger" can differ depending on interactions between different processes (Barrett, 2006). This theory is supported by numerous empirical evidence, including diverse brains scans and biological responses for supposedly the same emotion. However, so far this theory has not provided an exhaustive list of cognitive processes involved in creating complex emotions.

On the other hand, appraisal theories provide a list of cognitive appraisals responsible for creating emotions. In line with the constructionist theory, the EMA model (Marsella & Gratch, 2009) treats emotions as the results of continuous interactions between a basic set of emotional appraisals and other more complex cognitive processes. The appraisal mechanism checks several variables that are important for the formation of emotions, taking inputs from the external stimuli, bodily responses, and results of other cognitive processes. The appraisals are fast compared to other cognitive processes. They can happen iteratively or as other cognitive processes produce results. The emotions are created and coped with in an iterated cycle of appraisal, coping, and reappraisal. The appraisals in the EMA model include relevance, valence, intensity, future implications, blame/responsibility, power/coping potential, etc. The component process model (CPM) (Scherer, 2001; Scherer, 2009) investigates the relative speed of different appraisals. It contains four appraisal checks happening in a fixed order from fast to slow, namely: novelty and relevance, goal conduciveness, coping potentials and power, and adherence to personal and social standards. Although the appraisals happen in a fixed order, they constantly interact with other processes and may repeat iteratively. In the Iterative Reprocessing model (Cunningham, Dunfield, & Stillman, 2013), information is continuously processed in cycles to create emotions so that the boundary between emotional processes and non-emotional processes becomes blurred.

Systematically aligning these theories, especially the dual-process theory and the emotion theories, is beyond the scope of this paper. My theory on humor critically relies on the following propositions from emotion theories:

1. Emotion appraisals interact with other cognitive processes iteratively to produce subjective emotional experiences.

2. There is a fast appraisal for expectation violation or surprise, which executes without much conscious effort. In the CPM model, novelty (which is very similar to expectation violation and surprise, cf. Wessel et al., 2012) is the earliest appraisal. Neuroscience research (Holroyd

& Coles, 2002; Lieberman, 2003; Wessel et al., 2012) suggests the anterior cingulate cortex is primarily responsible for this function.

3. Many other cognitive processes, especially System 2, are slower than the surprise appraisal, so they can only produce results at later time.

3. A Synthetic Cognitive Theory of Humor

I propose that humor and the associated affect, mirth, is not an independent emotional or affective category. Instead, mirth is the result of a quick succession of several emotion appraisals and cognitive processes. In this section, I introduce a synthetic cognitive theory of humor. The following pun is used as an example to illustrate the theory:

Example 3.1. I asked if I was a gifted child, and dad said we wouldn't have paid for you. (Vaid et al., 2003)

The first emotion in the sequence is surprise, arising autonomously from System 1. The importance of surprise in humor comprehension has been noted by many humor theorists (Minsky, 1980; Veatch, 1998; Huron, 2008; McGraw & Warren, 2010; Hurley, Dennett, & Adams, 2013). Surprise is a slightly negative emotion, indicating errors in the expectation formed by System 1 (Holroyd & Coles, 2002). Neuroscience evidence suggests the anterior cingulate cortex (ACC) plays a central role in detecting errors we make, which are closely related to surprise. Holroyd and Coles (2002) reviewed multiple studies showing the ACC is most active at the initial stage of learning, which involves the most errors and the most surprises. Wessel et al. (2012) found the processing of novel and surprising stimuli share the same neural circuit within the ACC with the processing of error. Brain imaging studies (Mobbs et al., 2003; Watson, Matthews, & Allman, 2007) also found activities of the ACC during comprehension of humor. However, the ACC is also responsible for other cognitive functions, so the imaging results at the current resolution should be considered suggestive rather than conclusive.

Reading Example 3.1, it is clear that the surprise happens at the end of the sentence. As we read the phrase "gifted child", our associative reflex leads us to one particular meaning of the word "gifted". However, given our interpretation, the phrase "paid for you" at the end of the sentence does not make sense. This probably causes many readers to stop and think, or even go back and read the sentence again, more carefully this time.

Surprises may indicate cognitive errors or physical threats that we are unprepared for. As surprise is a fast appraisal, System 2 is quickly activated to deal with this unfamiliar situation (Lieberman, 2003). System 2 makes use of working memory and is correlated with dilated pupils, increased heart rate, and higher consumption of glucose (Gailliot & Baumeister, 2007; Kahneman, 2011). The mind is under stress to find out the source of the error and tries to correct it, in the hope that we can learn from the error and improve performance next time. If we realize the door knob is broken, we can find someone to fix it, or put up a sign to remind ourselves next time.

However, in the case of a joke, System 2 comes to the quick realization that the surprising stimulus is bogus: there is really no cognitive errors to correct and no threats (either physical threat or threats to values we hold dear) to escape from. We attribute the surprise signal to something we

know to exist, such as the two meanings of the same word, or to the intention of someone who played the pun on us. In Example 3.1, we realize there is little to learn because the joker tricked us into taking one meaning of the word, and of course we know both meanings of "gifted". We probably will not find it funny if we don't know both meanings because, for example, English is not our native language. In other jokes, we can often attribute the error to the stupidity and social inferiority of some story characters. That is, we dismiss any opportunities to change our understanding of the world or cope with an external threat. Therefore, we appraise the situation as irrelevant and not worth of further attention, and System 2 is deactivated. At the same time, a positive emotion, mirth, is created.

Appraisal theories lend support to the existence of the relevance check. Both the EMA model (Marsella & Gratch, 2009) and the CPM model (Scherer, 2001; Scherer, 2009) posit an relevance appraisal. In the CPM model, the relevance appraisal is a fast appraisal happening earlier than many other appraisals, but it is important to note the appraisal takes input from the slower System 2, which figures out the correct meaning of the sentence. That is why the dismissal is the third step of humor comprehension.

It is not always possible to dismiss the surprising stimulus as irrelevant. Not being able to do so negatively influences the perception of humor. Proulx et al. (2010) find that, after reading a parody story by Monty Python, those who perceive less threat to their values (and hence less need to re-affirm them) find the story to be funnier. The reason, according to the present theory, is that those who perceive threats to their values cannot dismiss the surprise as non-serious and irrelevant. The surprise-dismissal mechanism is similar to the incongruity-resolution (I-R) theory (discussed in Section 4.2). Differing from the I-R theory, my theory predicts that humor does not exist if the surprise is satisfactorily resolved and understood but not dismissed. This is demonstrated by surprises lead to genuine opportunities of learning, as in puzzles and riddles:

Example 3.2. What walks on four legs in the morning, two legs in the afternoon, three legs in the evening, and no legs at night?

People unaware of the answer are often surprised. After hearing the answer, the surprise is resolved, but the riddle is still not funny because people realize that the riddle served its purpose and perceive that they learned something from the answer.

But why is the quick dismissal and disengagement of System 2 processing funny? This can be understood as the trampoline effect, proposed by Huron (2008). As discussed earlier, System 2 consumes mental effort and strains the mind. Therefore, realizing the cognitive error is non-existent and disengaging System 2 lead to relief, a positive emotion. As surprise is a slightly negative emotion, when the system readjusts from negative to neutral or slightly positive, it can over-compensate and reinforce the subsequent outburst of positive emotion. Thus, a trampoline effect happens. Topolinski (2014) studies the fluency of humor processing and shows faster realization of the meaning of a joke leads to higher rating of its funniness. This may be interpreted as follows: the strength of surprise and its negativity is strongest at the moment surprise is created from appraisal. Over time, its strength diminishes. The strength of overcompensation correlates with that of surprise, and diminishes as System 2 takes time to make sense of the surprising stimulus.

Going back to Example 3.1, a significant portion of mirth is derived from the social inappropriateness, that the father implies his child is not worth much, and the self-deprecation of the speaker.

A proponent of superiority theory would argue this is the entirety of the joke. However, this stance is incompatible with other theories and cannot explain all types of jokes (See discussion in Section 4.1). The feeling of superiority by itself, such as beating an opponent in a difficult chess game, does not directly translate into humor. I contend that the feeling of superiority compound the effects of humor. There are two possible ways to unify the superiority theory with the current cognitive theory: One, the social inappropriateness and implied low social status is a reason for dismissing the surprising stimulus. Therefore, it is part of the surprise-reflection-dismissal-compensation process rather than an independent function. Two, the judgment of social appropriateness and social status leverages the surprise mechanism. Due to the trampoline effect (Huron, 2008), any positive emotion following the surprise and its dismissal can be amplified. Merely superiority may not be funny, but when it is amplified by the trampoline effect, it becomes funny. Ascertaining the exact mechanism may require further empirical studies.

We can summarize the four-step process of humor comprehension as: surprise, reflection, dismissal, and compensation. In the surprise step, an expectation created by the associative System 1 is violated, leading to surprise and the engagement of System 2. In the reflection step, System 2 makes sense of the surprising stimulus. In the dismissal step, we appraise the surprising stimulus, which now makes sense, as not worthy of further processing and dismiss it as irrelevant. In the compensation step, as the brain re-adjusts back from a slightly negative state to a positive state, a trampoline effect occurs, which can be compounded by other factors such as superiority or the pleasure of discovery. This dynamic process combines emotional appraisals and cognitive inferences to create the affective state known as mirth.

4. Existing Theories of Humor

In this section, I contrast my theory with existing theories on humor. As many theories of humor exist, and major theories tend to have more than one variant, I will limit myself to the most influential theories.

4.1 Superiority Theory

Example 4.1. How many Poles does it take to screw in a light bulb? Five. One to hold the light bulb, and four to turn the table he is standing on.[2] (Attardo & Raskin, 1991).

Superiority theory is one of the most popular and oldest theories of humor. It is evidently present in the example above. According to Attardo and Raskin (1991), this joke was popular when Polish immigrants were discriminated against in early American history. At different historic periods, the joke had many variants, poking fun at different people. Plato and Aristotle were both supporters of this theory.

Example 4.2. Two goldfish were in their tank. One turns to the other and says, "You man the guns, I'll drive." (Hurley, Dennett, & Adams, 2013)

Example 4.3. Photons have mass? I didn't even know they were Catholic. (Hurley, Dennett, & Adams, 2013)

2. It must be noted that the author does not support the racist view of this joke.

However, the superiority theory have difficulties explaining all kinds of jokes. The feeling of superiority by itself, such as beating an opponent in a difficult chess game, does not directly translate into humor. Puns, like those shown above, often do not have an obviously inferior individual and the superiority theory cannot explain them well. Admittedly, it is often possible to find the inferior individual using close reading. For example, one may argue that we feel superior than the illusionary goldfish, or the joker who did not understand the meaning of "mass". Nevertheless, I find it implausible that the human cognition would spend significant effort to find an inferior individual during online processing of a joke.

In the proposed theory, superiority compounds the effects of humor, but is not a defining feature of humor. Without clearly exhibited inferiority, humor is still possible. In those puns, the dismissal of surprise happens as soon as when we understood the double meaning of the word, and made sense of the entire passage. We do not postpone laughing until we clearly identify an inferior individual.

4.2 Incongruity-Resolution

Example 4.4. Everyone had so much fun diving from the tree into the swimming pool, we decided to put in a little water. (Binsted et al., 2006)

The incongruity-resolution (I-R) theory (e.g. Koestler, 1964; Suls, 1972) states that humor is created by the forming of an incongruity that is subsequently resolved. The I-R theory is also very popular and has many variants. Suls (1972) propose a two-stage process, starting with an expectation violation (i.e. an incongruity), followed by problem-solving activities that reconcile the incongruity. Minsky (1980) note a frame shift in puns, i.e. we realize one meaning of the word is wrong and shift to another meaning, where each meaning is represented by a frame. In Example 4.4, readers initially has an image of a swimming pool filled with water, as people for most of the time jump into water. As soon as they reach the end of the sentence, the readers realize the pool was empty, and switch to the frame of suicide. In another influential theory, Attardo and Raskin (1991) claim humor is created by the opposition of two scripts, and the two scripts must be activated at the same time in the reader's mind.

The contribution of this paper over existing I-R theories is two-fold: First, it clarifies the I-R theory and grounds it in modern cognitive science and neuroscience findings. My theory requires the surprise to be appraised as trivial and not an opportunity for learning, whereas the I-R theory only requires it to be "resolved". This helps to elucidate the theory, differentiate humor from riddles and puzzles, and explain the fact that jokes can be repeatedly funny (see Section 5.2). In Attardo and Raskin (1991)'s theory, it is not clear two opposing scripts can appear and how they can coexist. With the dual-process theory, it becomes clear that System 1 and System 2 can produce different interpretations, and only one is active after the punchline (Giora, 1991). Second, this paper reconciles the I-R theory with other competing theories, including the superiority theory, the release theory, and the recently proposed theory of mistake revelation (Hurley, Dennett, & Adams, 2013).

Another variant of the I-R theory has to do with morality and social norms. Veatch (1998) argue humor is the simultaneous recognition of a violation of moral norms and an interpretation that the situation is normal. As a recent update for Veatch, McGraw and Warren (2010) propose a two-step

process for humor comprehension. First, a violation of our expectation or social norms happens. This violation threatens to challenge our moral standards, humiliate us or shift our own beliefs about the world, which are all undesirable. However, we soon realize this violation is benign. For example, the person who was ridiculed is psychologically distant from ourselves. This leads to the feeling of mirth.

My theory is inspired by McGraw and Warren (2010), but provides a more concrete description for the word "benign". Our expectations are violated daily by events such as a bus running late or a computer being unresponsive. Most of these violations do not cause severe consequences, and could be called benign, but they are not funny. Another difference comes from the speed of the emotional appraisals involved. In the CPM model, the appraisal of social norms is the last of the four appraisals. Empirical evidence (Scherer, 2001; Scherer, 2009) suggests the cognitive process for determining if a behavior is approved by social norms is slower than processes associated with earlier checks. Therefore, I postulate this check usually comes later in humor comprehension. Its main effect is not to cause surprise or trigger System 2, but to compound the effect of the positive emotions produced by earlier processes.

4.3 Release and Relief

The release theory (Spencer, 1860; Freud, 1928) claims that humor is caused by release of wrongly mobilized psychic energy. The slightly different relief theory (Berlyne, 1972; Meyer, 2000) states that humor is caused by a release of nervous energy, and is employed in revealing suppressed desires such as sexual desires. In the same vein, Immanuel Kant claims:

> Laughter is an affectation arising from the sudden transformation of a strained expectation into nothing (Kant, 1790).

This type of theory of humor can appear overly philosophical and archaic, even bordering on mysticism. They certainly hint at some kind of mental effort, but descriptions such as "into nothing" and "psychic energy"seem too vague to be of any value. Nevertheless, when grounded in the dynamic and dual-process theory proposed in this paper, these claims can make sense.

As noted earlier, System 2 is engaged when surprised. The use of System 2 requires conscious effort and lead to pupil dilation and increased heart rate. Thus, the utilization of System 2 can be described as "mobilizing psychic energy". Kant's assertion of "nothing" coincides with the dismissal step, which stops System 2 from processing of the surprising stimulus further. It is therefore not unreasonable to describe the mobilized energy as "released". In this sense, my theory provides a refreshed, modern view of the relief and release theories of humor.

4.4 Revelation of Mistaken Beliefs

Minsky (1980) may be the first to suggest humor helps us learn to censor ridiculous thoughts. Developing the idea further, Hurley et al. (2013) propose a theory of humor and its evolutionary origin. According to this theory, humor is created when we realize one of our beliefs, which entered our mental space without our awareness, is wrong. Evolution has made recognition of mistaken beliefs fun, so as to encourage it. Thus, the sense of humor is evolutionarily adaptive. To Hurley et al, the basic form of humor is that I look for my glasses and find them on my ears.

My theory is influenced by and shares some similarities with Hurley et al.. Indeed, surprise and error are closely related and share the same neural circuitry (Wessel et al., 2012). However, we also differ in important ways. I do not argue the detection of errors, especially our own errors, is funny. Numerous studies on phenomena of confirmation bias, such as attitude polarization (Lord, Ross, & Lepper, 1979) and persistence of discredited beliefs (Ross & Anderson, 1982), clearly show that people are biased against recognizing their own mistakes. If a mechanism has been evolved to encourage us to recognize our own mistakes, it is not working very effectively.

My theory posit that the violated expectation is created by System 1, which often make mistakes, and quickly corrected and dismissed by System 2. Hence, we do not react to this violation of expectation overly negatively. In other words, humor comprehension does involve the realization that System 1 made an incorrect expectation, but this realization alone is not sufficient for humor.

Moreover, I do not argue humor is created by evolution as a separate mechanism for detecting error. It may be a side product of our ability to detect expectation violation, or evolved for its social function. Section 5.3 discusses social functions of humor.

My theory suggests a different basic form of humor, which happens when you raise a baby high in the air and lower him/her quickly. The baby is initially surprised, even slightly scared, by being raised. However, he/she then realizes there is no real danger. The giggle of the baby is the purest form of humor, not compounded by factors like superiority.

5. Interpreting Humor Phenomena

By now it is hopefully clear that the theory proposed in this paper makes a number of novel assertions compared to existing theories. However, the correctness of a theory must be tested with its consistency with empirical data. In this section, I will show that my theory of humor can explain several curious phenomena that are difficult to explain under previous frameworks. The most powerful evidences include the recently reported frustration smiles and the fact that jokes are often still funny after repetition. Furthermore, these evidences show that constituents of the theory, including the dual-process theory and the dismissal step, are necessary to explain the mechanism of humor.

5.1 Frustration Smiles

In a study that aimed to induce frustration, Hoque et al. (2012) from MIT put participants to solve impossible reCAPTCHAs, and recorded their facial expressions and actions with a webcam. After repeated failures at the reCAPTCHA, 90% of their participants produced genuine smiles during the experiment, even though the self reports contain only strong frustration. In a personal communication, Dr. Rosalind Picard explained that they found no existing theories to account for the cognitive mechanism underlying these "frustration smiles".

I believe the mechanism of frustration smile lies in emotion regulation (Gross, 2007; Gyurak, Gross, & Etkin, 2011). When we experience a negative emotion, we are motivated to cope with it by, for example, changing the external world or how we think about the world. One possible strategy is reappraisal, which actively modifies the results of emotional appraisals before the effects of emotion are fully appreciated (Gross, 2002). Instead of admitting that they failed at solving the reCAPTCHA, the participants in the frustration experiment can tell themselves that this is apparently

a bug in the program or a prank. The coping strategy of trivializing the stimulus serves the same function as the dismissal step of the humor process. As a result, the participants experience genuine mirth and exhibit genuine, or Duchenne, smiles.

Frustration smiles provide strong support for the proposed humor theory because it has a different mechanism. The experiment was never intended as a joke. The dismissal is an effect of active emotional reappraisal rather than a direct appraisal. The fact that the humor comprehension process can be triggered by different mechanisms with the same output strongly indicates its existence.

5.2 Jokes are Repeatedly Funny

If humor is a mechanism for us to discover cognitive errors and learn not to make them again (Minsky, 1980; Hurley, Dennett, & Adams, 2013), it is natural to expect the effect of humor should decrease as a joke is repeated. If humor exploits our expectation, it seems we should not create the same expectation again. However, several studies show that it is not always the case. Belch and Belch (1984) find low to medium levels of repetition actually increase the evaluation of humorous commercials, but high levels of repetition decrease their ratings. Zhang and Zinkhan (1991) find humor of TV commercials is unaffected by repetition. Using facial recognition software, Picard and colleagues find the second view of the same commercial can actually increase perceived joy (Collins, 2013).

Minsky (1980) hypothesized that some parts of our cognition may not learn very quickly. This paper provides a concrete grounding to that theory: As System 1 is associative and inflexible, it defaults to retrieve the most frequent solution, such as the most likely word sense based on the immediate context. If the audience of humor do not consciously suppress the automatic response of System 1, they will still experience surprise, which triggers the humor response. However, as the audience have seen the joke before, their System 2 can make sense of the surprising stimulus faster than when it was encountered the first time. This makes the second encounter funnier than the first due to the fluency effect (Topolinski, 2014). However, continued exposure of humorous stimuli will eventually train System 1 and the joke will be "worn out". Varying the message to violate the established expectation is shown to mitigate the wearing out (Belch & Belch, 1984).

5.3 Social Functions of Humor

Research shows that laughter and humor are significantly influenced by the social setting and serve social functions. Zhang and Zinkhan (1991) find jokes are funnier when someone else is present. Butcher and Whissell (1984) find the funniness rating of TV commercials increases as the number of viewers increases. Fridlund (1991) finds that even the presence of an imagined friendly companion can potentiate the smile action, regardless of the strength of the self-reported emotion. Meyer (2000) suggests that humor can create social bonding among those who laugh together as well as alienation for the butt of the joke. Multiple researchers (Miller, 2000; Bressler, Martin, & Balshine, 2006; Li et al., 2009) find that humor play an important role in how humans select their mates.

However, the question why humor should play such roles in human communication and reproduction has been left unexplained. The present theory suggests an asymmetry between the cognitive loads for joke crafting and joke understanding. The understanding of the joke partly utilizes the au-

tonomous and effortless System 1, and utilizes System 2 only briefly. In comparison, the joke teller need to walk a thin line. A good joke needs to be difficult enough so System 1 does not understand it, and easy enough so System 2 quickly understands it. The joke preferably should utilize other appraisals, such as superiority, to increase its funniness. Therefore, joke telling is a complex skill, and its mastery can indicate intelligence.

In addition, understanding a joke often involve many cognitive appraisals, including surprise, moral judgments, and social identities, so a joke becomes a condensed unit of information and a quick test for shared values and attitudes. Two people laughing at the same joke suggest many of their appraisals work similarly, therefore they share many values and attitudes. Not laughing often indicates otherwise. That's why laughing at one's boss's joke is widely believed to provide some career benefits. The role of humor in spouse selection can be explained as we look for people who are intelligent and share our values. Therefore, the social functions of humor provide one possible explanation for its evolutionary origin.

5.4 Individual Differences in Humor Appreciation

Whilst not well explored in the scientific literature (but see Forabosco & Ruch, 1994), anecdotal evidences indicate individual differences exist in humor comprehension. People favor different types of jokes. George Pierce Baker (1920) provided a differentiation of the so-called high and low comedy: "High comedy in contrast to low comedy rests then on thoughtful appreciation contrasted with unthinking, spontaneous laughter" (p. 236). Low comedy may include slapstick, farce, and jokes involving body parts, whereas high comedy may employ epistemic discovery such as those in puns. The dual-process theory is especially suitable for explaining this type of individual differences. As System 2 bears heavily on working memory capacity, individuals with lower working memory capacity may take more time to understand complex jokes and thus find them less funny due to the lack of processing fluency (Topolinski, 2014).

6. Implications for AI Systems

This paper mainly presents a cognitive theory on humor. However, I believe the theory in this paper can also provide some implications for AI researchers aiming to build general-purpose AI systems or systems dedicated to the creation and understanding of humor, as discussed below.

First, the present humor theory provides an account on the interplay between emotional appraisals and cognitive inferences in the comprehension of humor. It underscores the complexity of human affects such as humor, suspense and identification (Oatley, 1994), and the importance of building integrative AI architectures in order to model them. Phenomena like frustration smiles can only be understood in the context of interactions between multiple constituents of a large system. In particular, the human ability to monitor our own performance, detect errors, adapt, and self-regularize is central to this humor theory. Holroyd and Coles (2002) suggest error detection in the anterior cingulate cortex provides an important signal to reinforcement learning and built a neural network model to simulate error detection under an experimental condition. Extending existing systems (Cox & Ram, 1999; Schmill et al., 2011), Cox et al. (2011) propose a metacognitive ar-

chitecture designed around expectation and its violation. This theory re-affirms the need for similar cognitive architectures.

At least in principle, System 1 in the dual processes exhibits characteristics of a statistical system or an artificial neural network, whereas System 2 exhibits characteristics of a symbolic reasoning system. This suggests possible computational implementation for the two processes. O'Neill and Riedl (2014) model the recognition of suspense in narrative content with an activation-spreading network to represent the memory retrieval process and priming effects, and a symbolic planning system to represent problem solving.

Second, the construction of this theory suggests that cognitive theories and the development of AI systems can still inform each other. Granted, successful comedians are not cognitive scientists. It is possible to build computational systems that specializes in creating one type of jokes, and there are already a number of successful examples (cf. Binsted et al., 2006). However, the development of a general-purpose humor generator or humor understanding system can benefit from an understanding of the involved processes and their individual roles. Therefore, the author believes that it is necessary for cognitive science and AI to continue to learn from each other in deepening our understanding of the human intelligence, and intelligence in general.

7. Conclusions

Humor has been studied for centuries with many conflicting accounts for its working mechanism. In this paper, I provide a detailed cognitive theory of humor, which aims to ground humor comprehension in recent cognitive science research, unify different theories on humor, and explain various humor phenomena. This theory is built mainly on top of the dual-process theory and recent emotional theories. This theory describes the interactions between emotional appraisals, the automatic, associative, and effortless System 1, and the deliberate, flexible and effortful System 2.

According to the present theory, mirth is created by a quick succession of four emotional and cognitive phrases: surprise, reflection, dismissal, and compensation. In the surprise step, an expectation produced by the System 1 is violated, leading to surprise and the engagement of System 2. In the reflection step, System 2 makes sense of the surprising stimulus. In the dismissal step, the surprising stimulus is appraised as irrelevant and not worthy of further processing. In the compensation step, as the brain re-adjusts back from a slightly negative state to a positive state, a trampoline effect occurs, creating an outburst of positive emotions. The positive emotions can be compounded by other factors such as superiority or the pleasure of epistemic discovery. Converging evidences from brain imaging results, facial recognition studies, the effect of repetition on humor, and the recently reported frustration smiles support the validity of this theory.

Nevertheless, the theory should be considered as a working theory rather than as completely proven. A number of further substantiation and investigation are needed, especially on the compounding effect of superiority, and the mechanisms of the trampoline effect.

As an attempt for unifying existing theories, the current theory highlights the importance of studying complex affects, such as humor and suspense, in the context of interactions between cognitive processes and subsystems. Therefore, it provides motivations for the development of AI

systems that incorporate self-monitoring and error detection as a major building block. Building off recent advances of cognitive science, we may finally begin to crack the age-old mystery of humor.

Acknowledgements

I am grateful for discussions with Stacy Marsella, Brian Magerko, Rosalind Picard, Maarten Bos, Lew Lefton, and Pete Ludovice.

References

Attardo, S., & Raskin, V. (1991). Script theory revisi(it)ed: Joke similarity and joke representation model. *Humor*, *4*, 293–347.

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*, 184–206.

Bain, A. (1875). *The emotions and the will*. Longsman & Green. 3rd edition.

Baker, G. P. (1920). *The development of shakespeare as a dramatist*. New York: The Macmillan Company.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*, 28–58.

Barrett, L. F. (2011). Constructing emotion. *Psychological Topics*, *20*, 359–380.

Belch, G. E., & Belch, M. A. (1984). An investigation of the effects of repetition on cognitive and affective reactions to humorous and serious television commercials. In *Advances in consumer research volume*, Vol. 11, 4–10.

Berlyne, D. (1972). Humor and its kin. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor*, 43–60. New York: Academic.

Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., & O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, *21*, 59–69.

Bressler, E. R., Martin, R. A., & Balshine, S. (2006). Production and appreciation of humor as sexually selected traits. *Evolution and Human Behavior*, *27*, 121–130.

Butcher, J., & Whissell, C. (1984). Laughter as a function of audience size, sex of the audience, and segments of the short film 'duck soup'. *Perceptual and Motor Skills*, *59*, 949–950.

Collins, K. (2013). Rosalind picard on reading emotions by reading faces. Retrieved on Feb 23 2015 from http://www.wired.co.uk/news/archive/2013-10/17/rosalind-picard.

Cox, M. T., Oates, T., & Perlis, D. (2011). Toward an integrated metacognitive architecture. In P. Langley (Ed.), *Advances in cognitive systems, papers from the 2011 AAAI symposium*, 74–81. AAAI Press.

Cox, M. T., & Ram, A. (1999). Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence*, *112*, 1–55.

Cunningham, W., Dunfield, K., & Stillman, P. E. (2013). Emotional states from affective dynamics. *Emotion Review*, *5*, 344–355.

De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *Quarterly Journal of Experimental Psychology*, *59*, 1070–1100.

Evans, J. S. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.

Evans, J. S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

Forabosco, G., & Ruch, W. (1994). Sensation seeking, social attitudes and humor appreciation in italy. *Personality and Individual Differences*, *16*, 515–528.

Freud, S. (1928). Humor. *International Journal of Psycho-Analysis*, *9*, 1–6.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review*, *11*, 303–327.

Giora, R. (1991). On the cognitive aspects of the joke. *Journal of Pragmatics*, *16*, 465–485.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281–291.

Gross, J. J. (Ed.). (2007). *Handbook of emotion regulation*. New York, NY.

Gyurak, A., Gross, J. J., & Etkin, A. (2011). Explicit and implicit emotion regulation: A dual-process framework. *Cognition and Emotion*, *25*, 400–412.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

Hoque, M., McDuff, D., & Picard, R. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, *3*, 323–334.

Hullett, C. R. (2005). The impact of mood on persuasion: A meta-analysis. *Communication Research*, *32*, 423–442.

Hurley, M. M., Dennett, D. C., & Adams, R. B. (2013). *Inside jokes*. Cambridge, MA: MIT Press.

Huron, D. (2008). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kant, I. (1790). *Critique of judgement*. New York: Hafner Publishing Co. Republished in 1951.

Koestler, A. (1964). *The act of creation*. New York: MacMillan.

Li, N. P., Griskevicius, V., Durante, K. M., Jonason, P. K., Pasisz, D. J., & Aumer, K. (2009). An evolutionary perspective on humor: Sexual selection or interest indication? *Personality and Social Psychology Bulletin*, *35*, 923–936.

Lieberman, M. D. (2003). Reflexive and reflective judgment processes: A social cognitive neuroscience approach. In *Social judgments: Implicit and explicit processes*. Cambridge Univ. Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Marsella, S., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, *10*, 70–90.

McGraw, A., & Warren, C. (2010). Benign violations: making immoral behavior funny. *Psychological Science*, *21*, 1141–1149.

Meyer, J. C. (2000). Humor as a double-edged sword: Four functions of humor in communication. *Communication Theory*, *10*, 310–331.

Miller, G. F. (2000). *The mating mind: How sexual choice shaped the evolution of human nature*. New York: Doubleday.

Minsky, M. (1980). Jokes and the logic of the cognitive unconscious. A.I. Memo No. 603.

Mobbs, D., Greicius, M. D., Abdel-Azim, E., Menon, V., & Reiss, A. L. (2003). Humor modulates the mesolimbic reward centers. *Neuron*, *40*, 1041–1048.

Oatley, K. (1994). A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. *Poetics*, *23*, 53–74.

O'Neill, B., & Riedl, M. O. (2014). Dramatis: A computational model of suspense. In *Proceedings of the 28th AAAI conference on artificial intelligence*.

Proulx, T., Heine, S. J., & Vohs, K. D. (2010). When is the unfamiliar the uncanny? meaning affirmative after exposure to absurdist literature, humor, and art. *Personality and Social Psychology Bulletin*, *36*, 817–829.

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In *Judgment under uncertainty: Heuristics and biases*, 129–152.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research*, 92–120. Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*, 1307–1351.

Schmill, M., Anderson, M., Fults, S., Josyula, D., Oates, T., Perlis, D., Shahri, H., Wilson, S., & Wright, D. (2011). The metacognitive loop and reasoning about anomalies. In *Metareasoning: Thinking about thinking*, 183–198.

Schwarz, N. (2011). Feelings-as-information theory. In *Handbook of theories of social psychology*.

Spencer, H. (1860). The physiology of laughter. *Macmillan's Magazine*, *1*, 395–402.

Stanovich, K. (2011). *Rationality and the reflective mind*. Yale University Press.

Stanovich, K. E., & West, R. F. (2000). Individual difference in reasoning: implications for the rationality debate? *Behavioural and Brain Sciences*, *23*, 645–726.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons. In *Psychology of humor*, 81–99. Academic Press.

Topolinski, S. (2014). A processing fluency-account of funniness: running gags and spoiling punchlines. *Cognition and Emotion*, *28*, 811–820.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 239–278.

Vaid, J., Hull, R., Heredia, R., Gerkens, D., & Martinez, F. (2003). Getting a joke: The time course of meaning activation in verbal humor. *Journal of Pragmatics*, *39*, 1431–1449.

Veatch, T. C. (1998). A theory of humor. *Humor*, *11*, 161–215.

Watson, K. K., Matthews, B. J., & Allman, J. M. (2007). Brain activation during sight gags and language-dependent humor. *Cerebral Cortex*, *17*, 314–324.

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *The Journal of Neuroscience*, *32*, 7528–7537.

Zhang, Y., & Zinkhan, G. M. (1991). Humor in television advertising: the effects of repetition and social setting. In *Advances in consumer research*, 813–818.